individuals who indicate they have no chronic health condition. Results are compared using mean errors (ME), root mean squared errors (RMSE) and the proportion of values estimated within |0.05|. **RESULTS:** Using an age adjusted baseline, we found the additive (and multiplicative) methods underestimate the majority of HSUVs (ME:0.0781(0.0254); RMSE:0.1012(0.0651); 26%(56%) < |0.05|) while the minimum (and ADE) overestimate the majority of HSUVs (ME:–0.0995(–0.0695); RMSE:0.1214(0.0950); 20%(35%) < |0.05|). Although the simple linear model produced the most accurate results (ME:0.0001; RMSE:0.0598; 63% < |0.05|), there were some substantial errors with 20% of errors greater than the minimum important difference (|0.074|). When subgrouping by actual HSUV (range 0.350–0.917) we found the magnitude and direction of errors in the estimated HSUVs are driven by the actual HSUVs being estimated in addition to the technique used. In general the HSUVs estimated using an adjusted baseline were more accurate than those obtained using a baseline of perfect health. **CONCLUSIONS:** This study makes an important contribution to the evidence in this area as it is the first to compare the five different techniques in the same data set. While the simple linear model gave the most accurate results, the model requires validating in external data and additional research exploring an alternative model specification is warranted.

PMC34

ISSUES IN THE TRANSLATION AND LINGUISTIC VALIDATION OF CAREGIVER RATING SCALES REGARDING THE BEHAVIOR AND DEVELOPMENT OF CHILDREN AND YOUNG PEOPLE

Furtado T, Wild D
Oxford Outcomes Ltd, Oxford, Oxon, UK

OBJECTIVES: Caregiver rating scales, intended to evaluate the behaviour of children and young people, are frequently used in clinical trials involving youths. However, the translation and linguistic validation of such scales can be problematic due to the differing cultural markers of behaviour and development. This study aims to document the problems that can occur, with the hope of facilitating future studies and producing guidelines to avoid cultural compromises when such measures are developed. **METHODS:** Past Oxford Outcomes projects, which included the translation of Caregiver Reported Outcomes, were evaluated to identify problematic items. These included the Vineland-II and ABAS (behaviour development scales), ELDQOL (epilepsy and QoL scale) and WFI-RS (functional impairment rating scale) among others. **RESULTS:** Numerous cultural and linguistic issues became apparent, including the following:—Many examples of sports and activities were used in the documents, which required thorough cultural adaptation, e.g. types of games.—Logistical cultural differences were marked, e.g. questionnaires mentioned children's understanding of specific coins or traffic signals, which vary culturally.—Some documents involved markers for identifying speech development, such correct use of irregular verbs. These were problematic in other cultures and speech development specialists were required to find suitable alternatives.—More idiomatic expressions are used than in PROs developed for adults, e.g. "on the go"; these cause difficulties in translation.—Items surrounding activities such as housework or helping look after siblings are not equivalent in some cultures due to differing role expectations. **CONCLUSIONS:** The validation of caregiver reported outcomes through interviews with caregivers was particularly important with these scales to ensure cultural appropriateness in target languages. Physician or specialist input was sometimes required to find culturally relevant alternatives. When such measures are created, culturally specific markers of behaviour should be avoided if possible.

PMC35

RECOMMENDATIONS ABOUT TRANSLATIONS IN THE FINAL FDA GUIDANCE ON PRO MEASURES: WHAT HAS CHANGED AND WHAT HAS REMAINED

Conway K[1], Mear I[2]
[1]MAPI Research Trust, Lyon, France; [2]MAPI Institute, Lyon, France

OBJECTIVES: Almost four years were necessary to develop the final FDA guidance on the use of PRO measures in clinical trials. Our objective is to compare how the recommendations about translation and cultural adaptation evolved from the 2006 draft to the 2009 final guidance. **METHODS:** Both guidances were retrieved on the FDA website and analyzed. **RESULTS:** Structure and content were modified. Recommendations on translation and cultural adaptation were moved to another section within the Evaluating PRO Instruments Part: from "*IV.D. Modification of an existing instrument*" to "*III.G. PRO Instruments intended for specific populations*". As for the content, the text in the body of the final guidance is more concise compared to the draft. The novelty lies in the stipulation that the FDA will review the process used to translate/culturally adapt the instruments. As a consequence, an appendix (section VIII) was added in which the FDA explains which topics should be addressed in the documents provided to the FDA for review: description of process used, patient testing, rationale for decisions, copies of versions and evidence about validity. They are however key points which did not change: the need for providing evidence that content validity and other measurement properties are similar between all versions. **CONCLUSIONS:** The recommendations are more concise and precise, especially the expectations of the FDA. The FDA however does not indicate a preference for a specific translation methodology. Interestingly patient testing is clearly indicated as a key point of the process. The need for documenting all decisions is crucial and raises the question of developing standardized system of reporting to structure the evidence to be provided to the FDA. The last point of the Appendix is debatable as we anticipate that it might add a burden in term of costs to provide evidence about the psychometrics of all versions.

PMC36

SYSTEMATIC REVIEW OF THE RESPONSIVNESS OF SF-36 HEALTH SURVEY MEASURES TO EFFICACIOUS PHARMACEUTICAL THERAPIES IN WELL-CONTROLLED CLINICAL TRIALS

Ware JE, Frendl DM
University of Massachusetts, Worcester, MA, USA

OBJECTIVES: To determine how often SF-36 Health Survey measures respond to efficacious pharmaceutical treatment benefits in well-controlled clinical trials. **METHODS:** We conducted a systematic review of randomized, double-blind, placebo-controlled trials published in 124 journals in 1995 through 2009 documenting differences between treatment groups for primary medical endpoints and any of the SF-36 component summaries, or eight subscale scores. Concordance was defined in terms of agreement between primary clinical and SF-36 endpoints (both statistically significant or both non-significant). **RESULTS:** A review of 2,020 identified clinical trials using the SF-36 confirmed that 162 met study design criteria. For 133 of 162 trials (82.1%), results for primary clinical endpoints and SF-36 measures were concordant. Among the 107 trials achieving medical efficacy (primary endpoint), changes in one or more SF-36 measures were also significant, as hypothesized, for 88 (82.2%). Similar patterns were observed by therapeutic area; for example: rheumatology (29 of 30), neurology (16 of 25), cardiovascular (15 of 18), pulmonary (11 of 13), psychiatry (8 of 10), endocrine (7 of 9), and combined surgical specialties (9 of 9) studies demonstrated concordance. In addition to evaluating characteristics of published reports and scoring methods (subscales, summaries, utility scoring) this presentation will comment on priorities for future studies of patient-reported outcomes (PROs) in evaluations of pharmaceutical and other medical treatments. **CONCLUSIONS:** In support of their validity as PROs, changes in SF-36 measurements agree with primary endpoints in over 8 out of 10 well-controlled trials of pharmaceutical therapies published to date. In support of pharmaceuticals' efficacy, when a therapy positively impacted clinical endpoints, it also improved health related quality of life quality of life in over 8 out of 10 clinical trials published to date.

PMC37

DOES DATA COLLECTION FROM ONLINE COMMUNITIES RESULT IN BIASED RESPONSE?

Vaccarino AL[1], Sills TL[1], Bharmal M[2], Cascade E[3], Kalali AH[4], Evans KR[1]
[1]OCBN, Toronto, ON, Canada; [2]Quintiles, Rockville, MD, USA; [3]iGUARD Inc, Rockville, MD, USA; [4]Quintiles CNS Therapeutics, San Diego, CA, USA

OBJECTIVES: Although the ability to interact with patients in an on-line environment has expanded substantially over the past few years, many researchers are concerned that participants may not be representative from a medication experience perspective (i.e., biased towards complainers). The purpose of this study is to investigate patient responses on treatment satisfaction using a validated PRO measure, the Treatment Satisfaction Questionnaire for Medications (TSQM), collected through a survey of patients with depression from an on-line community. **METHODS:** A random sample of iGuard.org members treated with an antidepressant were invited to complete an online version of the TSQM, a widely used validated 14-item generic treatment satisfaction instrument. iGuard.org is an online patient community that provides a free medication monitoring service to patients. Non-parametric item response analyses were performed to determine the relationship between scores on individual items and total TSQM scores. **RESULTS:** Responses from 3641 patients were included in the analyses. TSQM Global Satisfaction scores ranged from 0—100 suggesting a broad spectrum of treatment satisfaction. Non-parametric Item Response analyses of raw scores revealed that individual items of the TSQM discriminated differences in patient satisfaction. That is, as total scores increased the probability of low scores on the individual items decreased and the probability of higher scores increased. As expected, patient satisfaction was related to reported side-effects, with those reporting side-effects experiencing lower satisfaction with medication than those without reported side-effects. **CONCLUSIONS:** The results from this analysis suggest that PRO survey data collected through a random sample of members of the on-line patient community iGuard.org can be representative of the spectrum of anticipated treatment satisfaction responses. Continuing to explore the potential of direct data capture from on-line patients will be important as researchers seek faster and cheaper alternatives to traditional physician-based recruitment.

PMC38

A COGNITIVE DEBRIEFING METHODOLOGY FOR ESTABLISHING EQUIVALENCE DURING E-PRO MIGRATION

Doyle S, Wild D
Oxford Outcomes Ltd, Oxford, Oxon, UK

BACKGROUND: Most outcomes instruments have been developed and validated as paper versions, but few have been migrated to electronic format. Migration to electronic delivery, without significantly altering format or text, qualifies as a minor modification not requiring a full validation (Coons et al. 2009). However, this does not mean that the two formats are perceived in the same way by patients. We aim here to describe a methodology successfully used to establish equivalence between paper and electronic PROs. **METHODS:** To demonstrate the equality of these different modes of data collection, we have used a combination of "think-aloud" and retrospective cognitive debriefing techniques, as well as usability testing. The debriefing exercise is designed to assess whether the electronic device changes the way respondents interpret the questions or response options. The usability testing assesses ease of use and identifies issues that may prohibit the use of the ePRO by the target

population. We typically recruit between 10–20 patients in which half the participants receive the ePRO first and the other half the paper version. Between administrations participants complete a distraction task. Interviews are recorded and a content analysis conducted to identify key issues. **RESULTS:** The mix of think-aloud and retrospective probing has worked well in a number of studies across disease areas to ensure equivalence, high usability, and no unforeseen issues unique to ePRO such as screen glare or difficulty holding a PDA device. Some patients have difficulty with the "think-aloud" approach and so the retrospective probing is a useful check against issues not spontaneously raised by the participant(s). **CONCLUSIONS:** Increased use of ePRO questionnaires necessitates a robust methodology for demonstrating equivalence during migration from paper versions. a mix of concurrent "think-aloud" and retrospective probing following completion of both PRO formats has shown to be a useful method for establishing validity of electronic outcome measures.

PMC39

EXAMINING ITEM RESPONSE PATTERNS OVER TIME IN A HEALTH PROFILE MEASURE USING US NATIONAL REPRESENTATIVE SAMPLES: A MULTI-FACET MODEL APPROACH

Gu NY

Pharmerit North America, LLC, Bethesda, MD, USA

OBJECTIVES: To examine item response patterns over time using the SF-12v2™ from a measurement perspective using US national representative samples. **METHODS:** Four panel data with two-year repeated measures on each respondent were extracted from the Medical Expenditure Panel Survey (MEPS). Respondents were included if they were ≥18 years, had completed SF-12v2™ and, had at least one of the top ten most prevalent health conditions identified using ICD-9-CM. Three-facet measurement model was used to parameterize time as a distinct facet in the model, in addition to person and item facets. Interactions between time and the twelve items were examined at each time point in all panels. Goodness-of-fit of the items to the model was examined in repeated measures as well as in point-in-time measures. INFIT mean-square (MnSq ≤ 1.40) was used as an item fit indicator. Cross-validations were conducted in each disease groups. **RESULTS:** Four panels were comparable in their distributions in health conditions, socio-demographics (mean ages were 52–53 years and, about 76–77% were white) and, sample sizes (2003–04, n = 2,124; 2004–05, n = 2,070; 2005–06, n = 2,148 and 2006–07, n = 2,329). Consistently in all panels, significant time and item interaction biases were found at time 1, especially on mental health items ($P < 0.01$). On the other hand, interaction biases between time and items at time 2 were not significant ($p > 0.05$). All items fit the model in repeated measures where time was parameterized as a facet (INFIT MnSq ≤ 1.40). The mental health item "*Have you felt calm and peaceful?*" consistently showed misfit in all point-in-time measures (INFIT MnSq > 1.40). Similar findings were noted in sub-samples. **CONCLUSIONS:** Findings from this study suggest consistent learned response patterns over time, especially the responses to mental health item, which give rise to the importance of inter-temporal health context in health measurement. Hence, cross-sectional health measures should be interpreted with caution.

PMC40

ITEM CALIBRATION OF A GENERIC ROLE FUNCTIONING ITEM BANK

Anatchkova M[1], Bjorner J[2]

[1]University of Massachusetts Medical School, Worcester, MA, USA; [2]National Research Centre for the Working Environment, Copenhagen, Denmark

OBJECTIVES: Role functioning (RF) is a key component of social well-being and thus an important outcome in health research. The aim of this study was to calibrate on a common metric newly developed items assessing the impact of health on RF. The items were developed based on review of the literature and focus group interviews and were found to be sufficiently unidimensional for item response theory applications. **METHODS:** Two thousand five hundred participants completed a battery of measures including 77 items in a RF bank, covering the impact of health on family, occupational and social role functioning. Each new item covered only one of the content areas. Items were evaluated for potential DIF by demographic variables (gender, age, and chronic condition) using a logistic regression approach. To estimate the item parameters for each domain on a common metric we used the generalized partial credit model. Item fit was evaluated using the $S\text{-}G^2$ index. Comparison of group mean bank scores of participants with different self-reported general health status and chronic conditions was used to test the external validity of the bank. **RESULTS:** After excluding items with DIF and poor fit the final item bank had a total of 64 items covering 4 general content areas of role functioning (family, social, occupational, generic). Slopes in the bank ranged between 0.96 and 4.51; the mean threshold range was –0.66 to –1.80. Item bank based scores were significantly different for participants with and without chronic conditions ($F(4, 2488) = 31.48$, $P < 0.0001$) and self-reported general health ($F(4, 2488) = 233.55$, $P < 0.0001$). **CONCLUSIONS:** An item bank assessing health impact on RF across 4 content areas has been successfully calibrated. Using computerized adaptive assessment, respondents will only need to answer items regarding relevant roles, while IRT score estimation still allows for scoring all respondents on the same common metric.

PMC41

PREEMPTING DIFFICULTIES IN LINGUISTIC VALIDATION, THE USE OF FACE VALIDATION TO CREATE MORE SOUND TRANSLATIONS

Gawlicki M[1], Handa M[2]

[1]Corporate Translations, Inc, East Hartford, CT, USA; [2]Corporate Translations, Inc, Chicago, IL, USA

OBJECTIVES: The process of linguistic validation is complex especially when working with a variety of languages in widely divergent cultural settings. The ability to clearly delineate concepts and synchronize wording within an instrument before the linguistic validation process begins not only significantly improves the original instrument, but also aids in optimizing its translatability, ensuring greater uniformity between multiple linguistic adaptations and saving time and resources along the way. This paper seeks to explain the benefits provided by the supplemental pre-translation process of face validation. **METHODS:** As part of a case study, face validated questionnaires were compared to the original homegrown versions of the corresponding instruments—questionnaires that were already psychometrically validated were not eligible. Changes that were made as a result of this analysis will be discussed in-depth to clarify difficulties that each issue would have created for the linguistic validation process had they not been corrected. a cost benefit-analysis was also conducted to confirm the value of this supplemental linguistic validation phase. **RESULTS:** While standard elements of the linguistic validation process, such as concept elaboration, international harmonization, survey research expert review, in-country clinician review and cognitive debriefing all assist greatly in creating a quality translation, none of their benefits are a substitute for face validation. Furthermore, cost-benefit analysis reveals that the preemption of linguistic or methodological issues prior to translation and the greater uniformity obtained amongst multiple translations created through face validation save time and money later on in the linguistic validation process, justifying the added up-front costs. **CONCLUSIONS:** As the case studies confirm, taking steps to maximize the translatability of a questionnaire prior to linguistic validation, through face validation in particular, is highly beneficial to the end-products and can also hasten overall project completion and improve the quality of all language versions of the instrument.

PMC42

TO WHAT EXTENT CAN TECHNOLOGY IMPROVE THE VALIDITY OF CLINROS?

Wild D[1], Langel K[2]

[1]Oxford Outcomes Ltd, Oxford, Oxon, UK; [2]CRF Health, Helsinki, Finland

OBJECTIVES: ClinROs are the most commonly observed endpoint in FDA approved product labels but few have been adequately scrutinized in terms of their suitability as endpoints. This study evaluates two widely used ClinROs (the Expanded Disability Status scale (EDSS), and the Hamilton Rating scale for Depression (HAM-D)) and provides an assessment on how migrating the measures onto an electronic platform might be able to improve their validity and reliability. **METHODS:** A literature review was conducted on both measures to evaluate the availability of information on their content validity and reliability and validity. An assessment was made on how the measures could be improved if they were to be migrated onto an electronic platform: **RESULTS:** The EDSS has shown varying results for validity and inter-rater reliability and it involves a complex scoring procedure. The migration of the EDSS onto an electronic format would enable an automated scoring system which could improve its validity. The HAM-D was found to be lacking in evidence of content validity and to have some complexity in the scoring system. Transferring the HAM-D onto an electronic platform could simplify the scoring system which could improve its validity. **CONCLUSIONS:** This study has highlighted some of the issues with validity and reliability of two widely used ClinROs. The migration of ClinROs to an electronic platform in addition to the ePRO migration cognitive debriefing and usability testing might go some way to improving the clarity of ClinROs which may go some way to improving the validity of the measures. It cannot however resolve all of the issues such as lack of content validity and its impact would vary widely according to the complexity of the ClinRO itself.

PMC43

DATA POOLING OF PATIENT-REPORTED OUTCOMES IN CLINICAL TRIALS: EVALUATION OF STATISTICAL TECHNIQUES FOR ASSESSING MEASUREMENT EQUIVALENCE

Nixon M

Quintiles, Bracknell, Berkshire, UK

OBJECTIVES: This analysis describes the development, application and comparison of three different approaches to evaluate measurement equivalence properties of a patient reported outcome (PRO) questionnaire applied to two treatment groups for gastroesophageal reflux disease (GERD). The data used in this analysis was obtained from an on-line patient community, iGuard.org. Patients using either of the two treatments were randomly invited to complete a measure of treatment satisfaction, the Treatment Satisfaction Questionnaire for Medication (TSQM). **METHODS:** Three statistical approaches were used to evaluate the measurement equivalence of the TSQM across the two patient populations: 1) Classical Test Theory (CTT) to assess the internal consistency of the TSQM items within each of the three factors using Cronbach's alpha; 2) Confirmatory Factor Analysis (CFA) using a special case of structural equation modelling (SEM); and 3) Item Response Theory (IRT)—based technique of Differential Item Functioning (DIF). **RESULTS:** All three statistical methods indicated measurement equivalence had been achieved across the two treatment populations for all the three domains of the TSQM. The effectiveness and global